Active Selection with Label Propagation for Minimizing Human Effort in Speaker Annotation of TV Shows

Budnik Mateusz, Johann Poignant, Laurent Besacier, Georges Quénot

Active Selection with Label Propagation for Minimizing Human Effort in Speaker Annotation of TV Shows

Mateusz Budnik[1,2]*, Johann Poignant*[1]*, Laurent Besacier*[1] *and Georges Quénot*[1,2]

[1]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2]CNRS, LIG, F-38000 Grenoble, France
first.lastname@imag.fr

Abstract

In this paper an approach minimizing the human involvement in the manual annotation of speakers is presented. At each iteration a selection strategy choses the most suitable speech track for manual annotation, which is then associated with all the tracks in the cluster that contains it. The study makes use of a system that propagates the speaker track labels. This is done using a agglomerative clustering with constraints. Several different unsupervised active learning selection strategies are evaluated.

Additionally, the presented approach can be used to efficiently generate sets of speech tracks for training biometric models. In this case both the length of the speech track for a given person and its purity are taken into consideration.

To evaluate the system the REPERE video corpus was used. Along with the speech tracks extracted from the videos, the optical character recognition system was adapted to extract names of potential speakers. This was then used as the 'cold start' for the selection method.

Index Terms: active learning, annotation propagation, clustering, speaker identification

1. Introduction

In this paper an efficient approach to speaker annotation is presented. Data annotation can be costly, especially given the amount of available information on the Internet and television. Additionally, it is still mostly unlabeled, which severely restricts its use. Active learning is one of the ways to address this problem by reducing the workload of the human annotator [1]. However, most active learning methods rely on a trained model to provide relevance or uncertainty scores [2, 3]. The drawback of this approach is that, e.g. when dealing with speaker identification, the list of classes (i.e. individual speakers) is not known beforehand and new classes may appear during the annotation process [4]. This can be seen in the case of video annotation where propagating the available labels can be as efficient as training a model [5]. Therefore, an approach that combines unsupervised active learning and label propagation is tested in this study.

An additional aspect analyzed in this paper is the possibility to create individual speaker corpora for training biometric models with the least amount of human effort involved. Biometric models could be then used for speaker identification in new videos or in unannotated audio data [6]. For evaluation, both purity and track duration of the available speech signal are taken into account.

This paper is organized as follows. Section 2 gives the overview of the system used for the experiments. The description of the optional modalities (written names and faces) are also included and followed by the presentation of the track selection strategies. Section 3 provides a description of the video corpus used in the subsequent experiments. Afterwards, a monomodal (based on speaker tracks only) evaluation is given. This is followed by the results of the two multimodal approaches: multimodal speaker annotation with the use of the overlaid names detection and speaker annotation using only faces labels. Section 4 concludes the paper and gives some perspectives.

2. System structure and base components

The structure of the system used in this study can be seen in Figure 1. First, speaker tracks are extracted from the videos and the distances between them are calculated. Optionally, face clustering and overlaid names can be obtained at this point as well.

The active learning cycle is then introduced. Here, based on the cluster structure and already available annotation (overlaid names or some previous available labels), a given selection strategy chooses an unlabeled sample for annotation. Once the new label is obtained, cluster recalculation and annotation propagation is done, which assigns a given label to the cluster containing the newly annotated track. During this process, some clusters may be combined or new ones created. This gives rise to a refined cluster structure, which serves as the basis for the next iteration of the active learning cycle.

2.1. Speaker diarization and generation of speech track distances

The speaker diarization system is straightforward and is done using conventional BIC-criterion [7]. After splitting the signal into acoustically homogeneous segments, the calculation of the similarity score matrix between each pair of speech tracks is done with the use of BIC with single full-covariance Gaussians. The distances are normalized to have values between 0 and 1.

2.2. Additional (optionnal) modalities

When dealing with complex data, like TV shows, more sources of information can be considered alongside speaker diarization. In this study two additional modalities were utilized: the written names that can be seen in an overlaid text whenever a given person is introduced, as well as faces extracted from the videos. These modalities are domain dependent and not always available even in video data, that is why in Figure 1 they are depicted as optional.

2.2.1. Overlaid names extraction

In order to be able to automatically extract the written names that appear in the video, an optical character recognition (OCR)

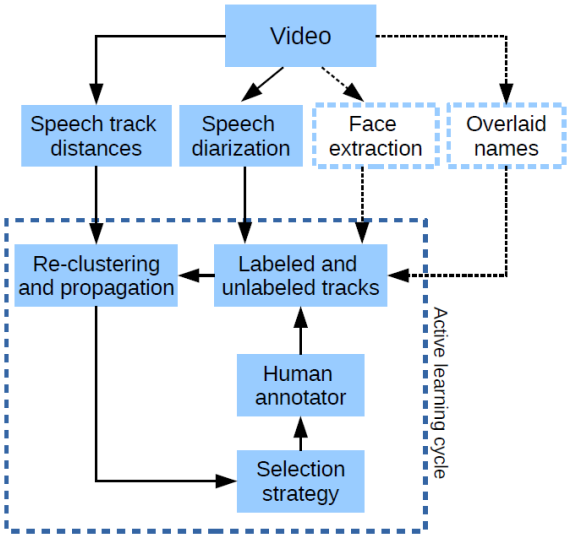

Figure 1: *System structure overview with optional multimodal parts: face extraction and overlaid text.*

system was used following the design proposed in [8]. In the context of this work, the obtained names serve as initial speaker labels, which are later expanded upon by the human annotator. Whenever a new guest is introduced in a given show, a text box usually appears containing his name. Not every text box contains a person's name however, some (especially in news broadcasts) display the name of the ongoing show, other the current news flash, etc. To address this issue, a list of names extracted from *Wikipedia* is used to verify the textual output.

Text detection and text recognition are the main steps in this method. For text detection a two step approach following [9] is adopted. The coarse detection is obtained through a Sobel filter and dilatation/erosion as in [10]. Additionally, to overcome the shortcomings of binarization, several binarized images are extracted from the same text, but temporally shifted. This is done to filter out false positive text boxes. For the text recognition part a publicly available OCR system from Google called `Tesseract` [11] was used.

2.2.2. *Face extraction*

The detail of the face extraction and clustering can be found in [12] and [13]. The faces are detected and tracked based on a the particle-filter framework using detector-based face tracker [14]. Three face detectors are used: frontal, half-profile and profile. With this information the most suitable faces are chosen based on a confidence score [15] and a local HOG descriptor is calculated on them [16]. After a dimensionality reduction step based on the LDML method [17], the Euclidean distance is calculated between every face track and a matrix of face track distances is obtained and then normalized between 0 and 1.

For the multimodal case (optional), in order to connect the face tracks that co-occur with speech tracks, additional features are used, such as lip activity, head size, etc. A multilayer perceptron is then trained on those features. The output of the model (with values in the range of 0 and 1) is then treated as the distance between speech and face tracks. This is used to produce initial multimodal clusters, i.e. clusters constructed from both the speech and face tracks.

2.3. Re-clustering and propagation of annotations

During the re-clustering step, some constraints are set to forbid merging the clusters (denoted as c) with different names (i.e. n) associated to them (i.e. $c(n)$). Note that clusters can have more than one name at this step. The agglomerative clustering algorithm is used for this purpose. The full list of constraints is as follows (based on [12]). The cases that allow two clusters c_1 and c_2 to be merged are:

- $c_1(\oslash) \cup c_2(\oslash) \Rightarrow c_{new}(\oslash)$
- $c_1(n_1) \cup c_2(\oslash) \Rightarrow c_{new}(n_1)$
- $c_1(n_1, n_2) \cup c_2(\oslash) \Rightarrow c_{new}(n_1, n_2)$
- $c_1(n_1, n_2) \cup c_2(n_1) \Rightarrow c_{new}(n_1)$

Two clusters cannot be merged when one of the situations seen below occurs:

- $c_1(n_1) \cup c_2(n_2) \Rightarrow \oslash$
- $c_1(n_1, n_3) \cup c_2(n_2) \Rightarrow \oslash$

Also, two different clusters with the same name assigned to them cannot have a co-occurring face track (in the multimodal cluster case), i.e. faces that appear at the same time in a video.

2.4. Active learning cycle

We propose to use an unsupervised approach to active learning, which is based mostly on the data structure, i.e. the (monomodal or multimodal) clusters, and the length of the speech tracks.

Several strategies were tested. Including benchmark approaches, which consist of random selection of speech tracks for labeling from the still unlabeled pool of tracks. Also, a chronological selection of of tracks, i.e. according to their order of appearance in the video, was tested.

A significantly better performance can be obtained with the approach that aims at choosing the longest track first, and subsequently, longest tracks without a manual or propagated label assigned to them. Algorithm 1 describes the method. Starting without any annotation, the longest track is selected and annotated. Afterwards the label is propagated to the corresponding cluster (after re-clustering). The next track is selected based on its duration and if it had a propagated label already. The algorithm stops when all tracks are either annotated manually or through propagation.

Data: A set of speech tracks $S = \{s_1, ..., s_N\}$
Result: A set of annotated speech tracks
$Ann = \{a_1, ..., a_N\} \subset S$
$Ann \leftarrow \oslash$;
Initialise a set with propagated annotation:
$Ann_{prop} \leftarrow \oslash$;
while $|S| \neq |Ann| + |Ann_{prop}|$ **do**
  $s_{temp} = max\ S \backslash (Ann \cup Ann_{prop})$;
  $Ann \leftarrow Ann \cup \{s_{temp}\}$;
  $Ann_{prop} \leftarrow\ propagate(Ann)$;
end
return Ann;

Algorithm 1: *Active learning cycle with longest track selection.*

3. Experiments

3.1. Data corpus

In the experiments, the REPERE corpus [18] was used. It consists of 205 videos, which sums up to the total length of around 40 hours. It contains recordings of 7 different TV shows from

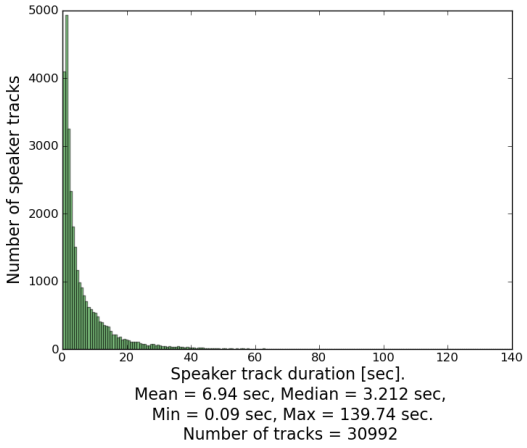

Figure 2: *Speaker track distribution and key statistical values.*

the French TV channels BFM TV and LCP. The videos are quite diverse in terms of their length (from about 3 minutes to half an hour) and the number of speakers present. Figure 2 shows the length distribution of the speech tracks extracted from the video corpus, along with some basic statistics.

3.2. Evaluation metrics to measure quality of speaker annotation

In this study the following evaluation metrics are used. First, the identification error rate (IER) evaluates the overall performance increase at every step of the active learning simulation.

Additionally, for assessing the quality of the annotation for every speaker (in order to train speaker biometrics models for instance), the purity is calculated. Then, every set of speaker track with purity score above 90% and above a given duration threshold is counted at each step of the experiment.

3.3. Experimental settings

In this study different selection strategies were evaluated. The experiment was a simulated active learning scenario where all the labels provided by human annotators are initially unknown and are revealed for a given speech track when the selection method selects it. At each step of the simulation (consisting of 20 steps in total) a single track is selected for labeling for every show as long as the new annotation is available. The whole experiment is repeated 10 times, at each time 80 % of the annotation per show is randomly selected, while the rest is not used in any way.

In the simulation, all the videos are processed in parallel. In terms of the computational time, the re-clustering and label propagation step in the case of the shortest video takes around 1 second, while for the longest it is around 40 seconds. Therefore, the computational time of a single step of the re-clustering is equal to the computation of the longest video. The computational time of the selection strategies is negligible.

3.4. Monomodal experiments

In this work two tasks were taken into account. On the one hand the efficiency of the annotation process is considered, in which the error reduction at each step is measured. Additionally, the ability to produce speaker corpora, which can later be used to train biometric models is also investigated.

In the case of the monomodal experiments, only the speech

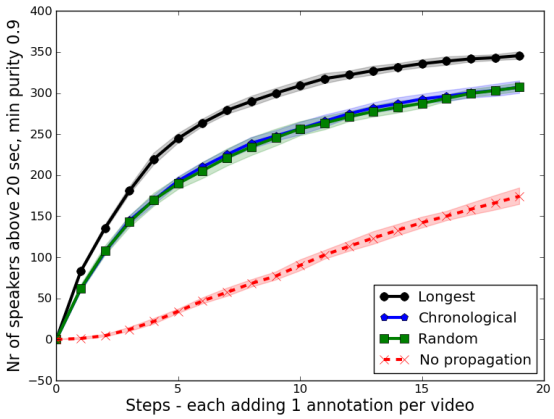

Figure 3: *Number of speakers with annotated tracks longer than 20 seconds (monomodal exp.).*

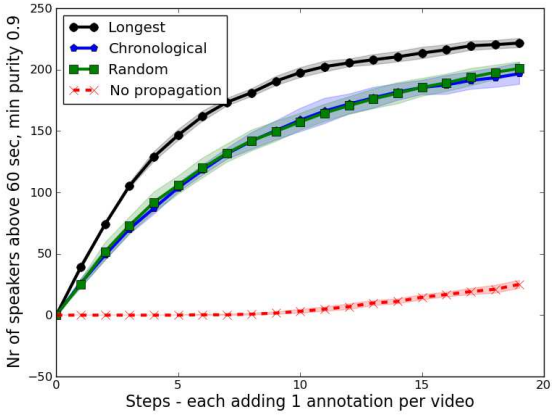

Figure 4: *Number of speakers with annotated tracks longer than 60 seconds (monomodal exp.).*

tracks extracted from the videos are used. At the beginning of the simulation, no annotation is available.

Figure 5 gives the identification error rate (IER) results for the monomodal speaker annotation task. In this and all subsequent plots the shaded area around the curves (with a corresponding color) is the standard deviation at each point. In addition to the strategies already mentioned, a strategy not making use of the label propagation is presented for reference. In this case, the selection of the tracks for annotation is done randomly.

Figures 3 and 4 show the number of obtained speaker corpora with purity score above 90% and with speech duration above 20 and 60 seconds respectively. For the 20 second condition, the proposed strategy works better than random at every step. Moreover, both approaches that make use of the annotation propagation are far better than the standard, no propagation method. The gap is even bigger when the 60 second condition is considered. Here the standard approach requires more than 9 steps (9 annotations per video) to produce any annotated speaker data meeting the criteria; and after 20 steps, it is still lower than when compared to the best strategy after a single step.

3.5. Multimodal experiments

In this section two multimodal experiments are presented. In the first one, the written names are used as the cold start for the active learning algorithm. In the second one, the head anno-

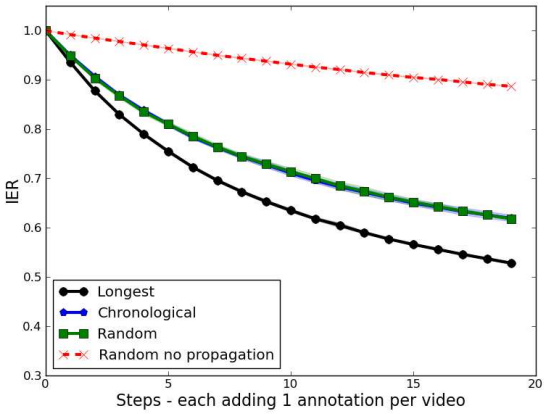

Figure 5: *Id. Error rate (IER) at every step of the active learning simulation (monomodal exp.).*

tation is used, instead of the speech annotation, but the results presented measure the speaker IER.

3.5.1. Overlaid text

In this experiment the co-occurring overlaid names were extracted from the video and used as an initial annotation for speakers. Afterward the annotation was further refined with the use of active learning. Figure 6 shows the result. When compared to the monomodal scenario this approach seems to be beneficial, also for the number of generated speaker corpora with the duration of 60 seconds or longer, which is equal to 315 after 10 steps for the longest track strategy against 190 for the corresponding approach without overlaid names.

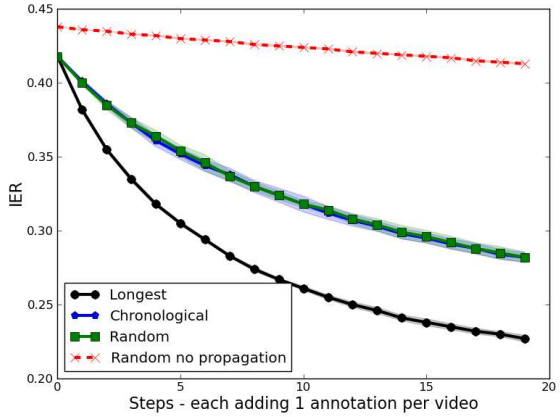

Figure 6: *Speaker annotation using overlaid text.*

3.5.2. Cross-modal effect

An additional experiment was done with the use of the head annotation only. In this scenario the human annotator would be asked to label faces rather then speech tracks. In this case, the speakers are annotated indirectly, through the use of multimodal clusters, which contain both the speech and face tracks. By labeling a face track, all the speech tracks in the cluster are also annotated. Figures 7 and 8 show the identification error rate measured on the speaker annotation exclusively with and without the overlaid names and with the random selection strategy. The results of annotation with the use of speaker tracks are provided in the corresponding plots for reference.

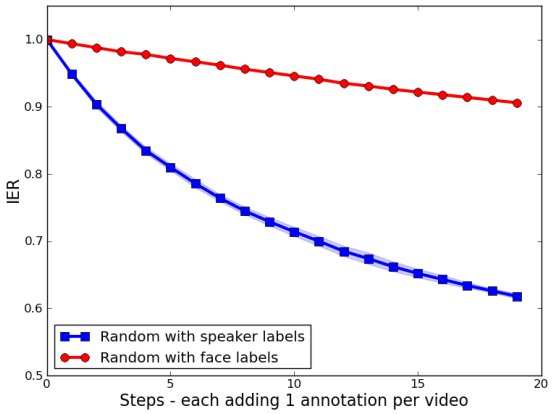

Figure 7: *Speaker annotation using face labels or speaker labels.*

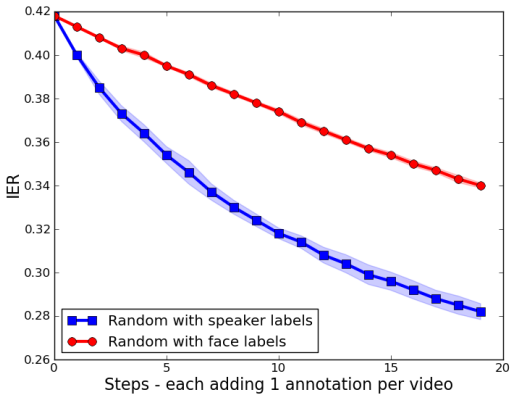

Figure 8: *Speaker annotation using face labels or speaker labels with overlaid text as an additional modality.*

The advantage of such an approach is that usually the process of face annotation is faster than speaker labeling. It is possible to present to the annotator a set of faces at the same time, while speech terms need to be heard one by one. The proposed approach makes it possible to produce annotations for two different modalities by presenting to the annotator just a single annotation task.

4. Conclusion and future work

This study presented an evaluation of an approach combining different selection strategies with the label propagation that can be adapted for efficient annotation of speaker in videos. This setting can provide a substantial reduction in the workload of the human annotator. When dealing with multimodal data, other sources of information can also be utilized, which can either improve the coverage (overlaid names) or simplify the task (face annotation) for the annotator. The future work include testing different clustering approaches, but also different selection strategies. An experiment with human participants will be considered with the help of on-line collaborative annotation platform with a graphical user interface.

5. Acknowledgements

This work was conducted as a part of the CHIST-ERA CAMOMILE project, which was funded by the ANR (Agence Nationale de la Recherche, France).

6. References

[1] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, pp. 55–66, 2010.

[2] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua, "Interactive video indexing with statistical active learning," *Multimedia, IEEE Transactions on*, vol. 14, no. 1, pp. 17–27, 2012.

[3] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," in *Advances in neural information processing systems*, pp. 892–900, 2010.

[4] W. Hu, W. Hu, N. Xie, and S. Maybank, "Unsupervised active learning based on hierarchical graph-theoretic clustering," *IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics*, vol. 39, no. 5, pp. 1147–1161, 2009.

[5] P. T. Pham, T. Tuytelaars, and M.-F. Moens, "Naming people in news videos with label propagation," *IEEE Multimedia*, vol. 18, no. 3, pp. 44–55, 2011.

[6] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 23–61, 2011.

[7] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, p. 8, Virginia, USA, 1998.

[8] J. Poignant, L. Besacier, G. Quénot, and F. Thollard, "From text detection in videos to person identification," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pp. 854–859, IEEE, 2012.

[9] M. Anthimopoulos, B. Gatos, and I. Pratikakis, "A two-stage scheme for text detection in video images," *Image and Vision Computing*, vol. 28, no. 9, pp. 1413–1426, 2010.

[10] C. Wolf, J.-M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, pp. 1037–1040, 2002.

[11] Google, "Tesseract–ocr." https://code.google.com/p/tesseract-ocr/, 2008.

[12] J. Poignant, H. Bredin, L. Besacier, G. Quénot, and C. Barras, "Towards, a better integration of written names for unsupervised speakers identification in videos," in *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM)*, pp. 84–89, 2013.

[13] J. Poignant, *Identification non-supervisée de personnes dans les flux télévisés*. PhD thesis, Université de Grenoble, 2013.

[14] M. Bauml, K. Bernardin, M. Fisher, H. K. Ekenel, and R. Stiefelhagen, "Multi-pose face recognition for person retrieval in camera networks," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pp. 441–447, 2010.

[15] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy–automatic naming of characters in tv video," in *Proceedings of the 17th British Machine Vision Conference*, 2006.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[17] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Face recognition from caption-based supervision," *International Journal of Computer Vision*, vol. 96, no. 1, pp. 64–82, 2012.

[18] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition.," in *LREC*, pp. 1102–1107, 2012.